Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Technology 6 (2012) 460 – 468

Procedia Technology

2nd International Conference on Communication, Computing & Security [ICCCS-2012]

Hierarchy classification for Data Warehouse: A Survey

Kanika Talwar[a], Anjana Gosain[b*]

[a,b]*Guru Gobind Singh Indraprastha University,Sector-16-C Dwarka,New Delhi and 110075, India*

Abstract

In data warehouse systems, the hierarchies play a key role in processing and monitoring information. These hierarchies dynamically analyze huge volumes of historical data in data warehouses at various granularity levels using OLAP operations like roll-up and drill-down. Through these operations we can get summarized as well as detailed data which aids in analysis as well as decision making process. Several authors have defined hierarchies deriving from real-world applications in order to represent broad range of business scenarios. But there is a need to properly categorize dimension hierarchies so as to adequately model them during evolution. In this paper we have provided a comprehensive comparison of different categories of hierarchies proposed by various researchers based on certain parameters.

Keywords: Data warehouse systems; Multi-dimensional schema; Dimension hierarchies.

1. Introduction

A data warehouse is the central repository which stores information, required by the top level managers and economic analysts of an organization for decision making. It is a subject-oriented, integrated, time-variant and non-volatile collection of data [13].The data warehouse's structure is generally represented with the help of the multi-dimensional schema which formulates information as facts and dimensions. Fact is a numeric value of a normally additive nature [14]. The fact expresses the focus of analysis in an enterprise and is illustrated through

* Corresponding author: Tel.: +91-9811055716
E-mail addresses: anjana_gosain@hotmail.com (Anjana Gosain), kanika.ncce@gmail.com (Kanika Talwar)

doi:10.1016/j.protcy.2012.10.055

a set of attributes called measures. The dimension allows the user to explore the measures from various perspectives of analysis. The hierarchies defined on various dimension attributes are majorly essential as the consequent data analysis might be addressed by these. The dimension hierarchies are used in a data warehouse to view data at different levels of granularity. These hierarchies allow the user to begin with a general view of data and achieve a detailed view with the drill-down operation. On the other hand, the roll-up operations transform detailed measures into summarized data.

From a technical point of view, a hierarchy is defined as a set of binary relationships existing between dimension levels, where a dimension level participating in a hierarchy is called hierarchical level or simply level. Given two consecutive levels of a hierarchy, the higher level is called parent and the lower level is called child [3]. A hierarchy represents some organizational, geographic, or other type of structure that is important for analysis [15]. Thus, supporting different kinds of hierarchies in the dimensional data, and allowing more flexibility in defining the hierarchies can enable a wider range of business scenarios to be modelled [16].

In the literature, several authors [1,2,3,4,5,6,7,8,9,10,11,12] have proposed different types of hierarchies which take part in analyzing information. In [2], the authors have derived dimension hierarchies from generalization and aggregation hierarchies and structured them as UML metamodels. In [1,7,10], dimension hierarchies are defined for OLAP cube. The goal of this paper is to present a survey of efforts done by various researchers for the categorization of hierarchies in context of DW. The paper is organized as follows. Section 2 discusses various kinds of hierarchies including their notations as proposed by various authors. Section 3 presents the comparative analysis of the related work in a tabular manner based on certain parameters. Lastly, Section 4 gives conclusion and future perspectives.

2. State of the Art

The below sub-sections discuss about various hierarchy categorization.

2.1. Niemi et. al. [2001]

In this paper, the authors [1] have classified various data hierarchies (listed below in table 1) in accordance to their generality for the OLAP cube. The most general is the acyclic digraph and the strictest is the balanced and non-ragged tree.

Table 1. Dimension hierarchies

Hierarchy	Description
Acyclic digraph	This hierarchy class allows “direct shortcut" and also has redundant aggregation paths.
Transitive anti-closed digraph	In this there are no “direct shortcuts” and no redundant aggregation paths are possible.
Tree	In this hierarchy, unique aggregation paths are guaranteed.
Unbalanced but non-ragged	The available levels of aggregation are not equal.
Balanced but ragged	The available levels of aggregation are not equal.
Balanced and non-ragged	In this equal levels of aggregation are available.

2.2. Akoka et. al. [2001]

This paper [2] concentrates on the definition of multidimensional hierarchies (described in table 2). The authors present and illustrate certain mapping rules for defining multidimensional hierarchies from UML schemas, based on aggregation and generalization hierarchies.

Table 2. Dimension hierarchies

Hierarchy	Description
An overlapping/ disjoint specialization hierarchy	This specialization is tackled by rule R5 i.e. For each level i of specialization of a class C, a class named Type-C-i is created.
Multiple inheritance hierarchy	Rule 10 tackles multiple inheritances, i.e. If a class C is a child of both G1 and G2 generalizations.
Multi-level generalization hierarchy	Rule 11 handles this hierarchy, i.e. If a class C is a child of a generalization G1 and parent of a generalization G2, a second aggregation level and a second class are created.
1-N aggregation	Rule R21 deals with the most classical case of aggregation, i.e. the 1-N aggregation transformed into a single dimension link between two dimensions.
1-N aggregation path	This is also handled by rule R21.
1-1 aggregation	This aggregation is handled by rule R21.
M-N aggregation	Rule R20 transforms the M-N aggregations into a plural dimension link in the logical level.

2.3. Malinowski, Zimányi [2004]

In this paper, the authors [3] present a conceptual classification of OLAP hierarchies and propose graphical notations for them based on the ER model which is summarized in table 3.

Table 3. Dimension hierarchies

Hierarchy	Description
Simple	This hierarchy can be represented as tree. They use only one criterion for analysis.
Symmetric	This has at the schema level only one path where all levels are mandatory i.e. branches of tree have same length.
Asymmetric	This has at the schema level only one path where all levels are not mandatory i.e. branches of tree have different length.
Generalized	This hierarchy can contain multiple exclusive paths sharing some levels. All have same analysis criterion.
Non-strict	This has at least one many-to-many cardinality i.e. a child member may have more than one parent member.
Symmetric Non- strict	This has at least one many-to-many cardinality, also all levels are mandatory.
Multiple	This hierarchy contains multiple non-exclusive simple hierarchies sharing some levels, accounting for same analysis criterion. Two types: (a) Multiple Inclusive (b) Multiple Alternate hierarchy.
Strict	In this, all cardinalities are one-to-many.
Parallel	Parallel hierarchies arise when there are multiple hierarchies, accounting for different analysis criteria. Two types: (a) Parallel independent (b) Parallel dependent hierarchy.

2.4. Lin Yuan [2006]

In this paper, the authors [4] have defined hierarchies on OLAP data cube dimensions. Here an iteration-based strategy is presented that aims to speed up rollup operation on recursive hierarchies in OLAP. The authors have discussed only two hierarchies which are as follows in table 4.

Table 4. Dimension hierarchies

Hierarchy	Description
Regular hierarchy	This hierarchy is balanced and is defined by an ordered list of levels.
Recursive hierarchy	This hierarchy is unbalanced and is defined by a self-referenced level.

2.5 Mansmann et. al. [2006]

In this paper, the authors [5] have classified various hierarchies (discussed in table 5) along the dimension based on granularity level. A framework is proposed for modeling complex hierarchical dimensions.

Table 5. Dimension hierarchies

Hierarchy	Description
Non-strict	A dimension allows many-to-many relationships between its levels.
Strict	A dimension with only one outgoing rolls-up relationship per entity
Non-hierarchy	A dimension with a single granularity
Multiple	A single dimension may have several aggregation paths.
Heterogeneous	Each subclass has its own attributes and aggregation levels resulting in heterogeneous subtrees in the data hierarchy.
Non-covering	In this hierarchy, some intermediate levels are skipped i.e. not covered.
Non-onto	It is a kind of Strict hierarchy that allows childless non-bottom nodes.
Mixed-granularity	This hierarchy contains mixed granularity which results in unbalanced tree.

2.6 Sergio Lujan-Mora et. al. [2006]

This paper [6] contains various classification hierarchies defined on dimension attributes. The authors have used a common way of representing and considering dimensions with their classification hierarchies by means of Directed Acyclic Graphs (DAG).The hierarchies are discussed in table 6.

Table 6. Dimension hierarchies

Hierarchy	Description
Multiple classification hierarchies	In this a dimension attribute may also be aggregated to more than one hierarchy.
Strictness	Strictness is a concept which means that an object of a lower level of a hierarchy belongs to only one of a higher level.
Alternative path hierarchies	This hierarchy contains two different paths that converge into the same hierarchy level.
Completeness	Completeness concept means that all members belong to one higher-class object and that object consists of those members only.

2.7 Vinnik, S.et. al. [2006]

The paper [7] describes hierarchy within an OLAP cube in following categories listed in table 7. The authors present a navigation framework for advanced exploration and analysis of multidimensional data in a data warehouse context.

Table 7. Dimension hierarchies

Hierarchy	Description
Simple	A non-hierarchical dimension i.e. no roll-ups or drill-down relationships exist.
Single	In this hierarchy, just a single decomposition path exists.
Multiple	This hierarchy exists when a dimension is subdivided in multiple ways.
Composite	This hierarchy exists when a dimension unites heterogeneous members from multiple relations in a single super class.
Mixed-Level	This hierarchy has mixed hierarchy levels. i.e. the entities from upper hierarchy levels do not merely serve for aggregating but also participate as end-entities in the fact table.

2.8 Mansmann, S. et. al. [2007]

In this paper [8] the authors have referred the hierarchies in the form of nodes. Also roles and types are defined for the nodes which together define its query behavior. The two main types of hierarchies discussed are: Schema and Data hierarchies which are discussed in table 8.

Table 8. Dimension hierarchies

Hierarchy	Description
Non-hierarchical node	It is a bottom-level category with the values of the finest granularity.
Single hierarchy	It is a non-abstract node with a single child category.
Multiple hierarchies	It is an abstract node with multiple child categories for the respective aggregation paths.
Super-class	It is an abstract parent of a single or multiple categories.

2.9 Anna Rozeva [2007]

In this paper [9], the author has discussed various hierarchies in terms of dependencies that should hold true. These dependencies can be forced on relation for ensuring correctness of aggregations along the levels. The complete sets of hierarchies are given in table 9.

Table 9. Dimension hierarchies

Hierarchy	Description
Transitive anti-closure dependency	For multiple hierarchies a constraint that removes multiple paths from a parent to child level is posed by the transitive anti-closure dependency.
Non-raggedness dependency	The non-raggedness dependency ensures that no levels can be bypassed in roll-up paths but still it doesn't provide for correct aggregations if hierarchy is unbalanced
Balance dependency	It states that leaf nodes are at the same level.
Functional dependency	It forces hierarchy to a tree. It enforces the rule that each child has a single parent.

2.10 Scholl et. al. [2007]

In this paper, the authors [10] have presented a case study of dimension hierarchies organized in form of OLAP cube. The major classification is between simple and multiple hierarchies. It also includes schema and instance normalization techniques for mapping conceptual design to logical design. It consists of 16 hierarchies listed in table 10 as follows: -

Table 10. Dimension hierarchies

Hierarchy	Description
Multiple	Multiple hierarchies exist whenever more than one hierarchy is defined within a dimension.
Non-strict	This has at least one many-to-many relationship between its categories
Strict	In strict hierarchy, each category has at most one outgoing roll-up relationship i.e. one-to-one cardinality.
Weighted non-strict	This hierarchy restores the summarizability by enforcing to specify each element's degree of belonging to each of its parent elements.
Non-covering	In this exclusive paths are obtained by allowing the roll-up relationships to be partial.
Simple	In this, a single hierarchy is defined upon the dimension and there is one analysis criterion. Two types: (a) Homogeneous (b) Heterogeneous
Symmetric	This hierarchy is a simple hierarchy in which all levels in the schema are mandatory thus forming a balanced tree.
Asymmetric	This simple hierarchy allows childless members in non-bottom categories. Also named as Non-Onto hierarchy.
Non-hierarchy	Here dimension consists of a single category i.e. not involved in any incoming or outgoing roll-up relationship.
Generalized	It contains categories that can be represented by a generalization relationship
Con-covering	This hierarchy preserves its schema and is passed over to the instance normalization phase.
Generalization	This hierarchy uses super classes for uniting multiple categories to treat their members as one category.
Specialization	This hierarchy uses subclasses as roll-up categories of the super class category. Two types: (a)Disjoint specialization (b)Overlapping specialization
Mixed	This is a special case of a generalized hierarchy, in which a roll-up relationship exists between the subclass categories of the same super class.
Multiple alternative	These hierarchies are non-exclusive aggregation paths with at least one shared level in the dimension schema. (a) Patterned hierarchy.
Parallel	These hierarchies in a dimension account for different analysis criteria but have multiple exclusive paths. Two types: (a)Parallel independent (b) Parallel dependent

2.11 S. Banerjee et. al. [2009]

In [11], authors concentrated on various approaches of schema evolution. Also, they have discussed about the core and additional features in terms of hierarchy (Table 11).

Table 11. Dimension hierarchies

Hierarchy	Description
Multiple	These hierarchies exist when a dimension can have multiple paths to roll-up or drill-down information.
Non-strict	This exists when a dimension can have many-to-many relationships.
Non-onto or missing data	This hierarchy exists when lower level in a dimension can exist without a corresponding data in the higher level to roll-up to.
Non-covering	Also known as unbalanced hierarchies or ragged dimensions. A ragged dimension is a dimension with at least one member whose logical parent is not in the level immediately above the member.

2.12. Zaker et. al. [2009]

In this paper, the authors [12] have discussed different hierarchies in terms of tree structures which are listed below in table 12.

Table 12. Dimension hierarchies

Hierarchy	Description
Balanced tree structure	In this structure, hierarchy has a consistent number of levels and each level can be named. Each child has one parent at the level immediately above it.
Variable depth tree structure	In this structure, the number of levels is inconsistent and each level cannot be named.
Ragged tree structure	This hierarchy has a maximum number of levels, each of which can be named and each child can have a parent at any level (not necessarily immediately above it).
Complex tree structure	In this hierarchy, a child may have multiple parents.
Multiple tree structures	This hierarchy is for the same leaf node [17]

3. Comparative Study

We have analyzed the various research works on several parameters and presented their comparison below in the table 13. In the table the symbol '✖' indicates that the corresponding parameter does not exist and symbol '√' indicates its existence in the related paper.

Table 13. Comparison of various research works

Parameters / Authors	Basic Structure	Generalization Relation-ship	Specialization Relation-ship	Dependency Support	Type of Design Considered	Approach used to represent hierarchies	One to Many Cardinality	Many to Many Cardinality
Niemi (2001) [1]	Tree + Graph	✖	✖	√	Logical design	Relational database theory	√	✖
Akoka (2001) [2]	UML + Multidimensional Meta models	√	√	✖	Conceptual + Logical design	UML Schema concepts+ Unified multi dimensional model	√	√
Malinowski (2004) [3]	Tree + Graph	√	✖	✖	Conceptual design	Graphical notations based on ER model	√	√
Lin (2006) [4]	Tree	✖	✖	✖	Conceptual design	Based on Object relational DBMS	√	✖
Mansmann (2006) [5]	Tree	✖	✖	✖	Conceptual design	Multi-dimensional ER model with EER notation	√	√
Sergio Lujan-Mora (2006) [6]	Directed Acyclic Graph	✖	✖	✖	Conceptual design	Extension of Unified Modeling Language	√	✖
Vinnik, S (2006) [7]	Decomposition Tree	✖	✖	✖	Logical design	Entity Relationship model	√	✖
Mansmann (2007) [8]	Tree	✖	✖	✖	Logical design	Entity Relationship model	√	✖
Anna Rozeva (2007) [9]	Tree + Graph	✖	✖	√	Logical design	Relational model	√	√
Scholl (2007) [10]	Tree + Graph	√	√	✖	Conceptual + Logical design	Multi-dimensional ER model with EER notation	√	√
S. Banerjee (2009) [11]	Tree + Graph	✖	✖	✖	Conceptual + Logical design	Graphical notations based on ER model	√	√
M. Zaker (2009) [12]	Tree	✖	✖	✖	Logical design	Relational model	√	√

4. Conclusion and future work

A hierarchy is a very important aspect of data analysis in a data warehouse, which defines the relationships between attributes of the dimension [18]. This paper mainly focuses on the hierarchy categorization done by various researchers and presents a comprehensive survey of dimension hierarchies based on several parameters. Various authors have proposed dimension hierarchy types illustrating relationships like aggregation, generalization or specialization and cardinality like one-to-many or many-to-many. This survey distinguishes complex hierarchies like non-strict, generalized, parallel, multiple [3] etc. from the simpler ones. Also we can clearly identify the method used for modeling hierarchies during evolution and for which hierarchy we need to solve aggregation issues. Our future scope includes handling hierarchies in case of data warehouse evolution and also to propose evolution operators to handle complex hierarchies in schema evolution.

References

[1] Niemi, T., Nummenmaa, J., & Thanisch, P. (2001). Logical multidimensional database design for ragged and unbalanced aggregation. DMDW'2001: Proceedings of 3rd International Workshop on Design and Management of Data Warehouses (pp. 7.1-7.8). Interlaken, Switzerland.

[2] J. Akoka, I. Comyn-Wattiau, N. Prat, Dimension hierarchies design from UML generalizations and aggregations, in: Proc. of the 20th Int. Conf. on Conceptual Modeling, 2001, pp. 442–445.

[3] E. Malinowski, E. Zimanyi, OLAP hierarchies: A conceptual perspective, in: Proc. of the 16th Int. Conf. on Advanced Information Systems Engineering, 2004, pp. 477–491.

[4] Lin Yuan and Hengming Zou, "Fast Rollup on Recursive Hierarchy in OLAP": Proceedings of 2006 IEEE/ACM/WIC, International Conference on Web Intelligence.

[5] Mansmann, S., & Scholl, M. H. (2006). Extending visual OLAP for handling irregular dimensional hierarchies. DaWaK'06: Proceedings of 8th International Conference on Data Warehousing and Knowledge Discovery (pp. 95-105). Krakow, Poland.

[6] Sergio Lujan-Mora , Juan Trujillo , Il-Yeol Song, A UML profile for multidimensional modeling in data warehouses: Data & Knowledge Engineering 59 (2006) 725–769.

[7] Vinnik, S., & Mansmann, F. (2006). From analysis to interactive exploration: Building visual hierarchies from OLAP cubes. EDBT 2006: Proceedings of 10th International Conference on Extending Database Technology (pp. 496- 514). Munich, Germany.

[8] Mansmann, S., & Scholl, M. H. (2007). Exploring OLAP aggregates with hierarchical visualization techniques. SAC 2007: Proceedings of 22nd Annual ACM Symposium on Applied Computing, Multimedia, & Visualization Track (pp. 1067-1073). Seoul.

[9] Anna Rozeva , Dimensional Hierarchies – Implementation in Data Warehouse Logical Schema Design: International Conference on Computer Systems and Technologies - CompSysTech'07

[10] Svetlana Mansmann, Marc H. Scholl (2007). Empowering the OLAP Technology to Support Complex Dimension Hierarchies: International Journal of Data Warehousing & Mining, 3(4), 31- 50.

[11] S.Banerjee, K.C.Davis, Modeling Data Warehouse Schema Evolution over Extended Hierarchy Semantics, S.Spaccapietra et.al (EDs): Journal on Data Semantics XIII, LNCS 5530, pp.72- 96,2009.@Springer- Verlag Berlin Heidelberg 2009.

[12] M. Zaker, S. Amnuaisuk, S. Haw, "Hierarchical Denormalizing: A Possibility to Optimize the Data Warehouse Design" International Journal of Computers, Issue 1, Volume 3, 2009.

[13] Inmon, W.: Building the Data Warehouse, pp 23 (1991).

[14] Multi-Dimensional Modeling with BI, Version 1.0, May 16, 2006: ©2000 SAP AG and SAP America, Inc.

[15] E. Malinowski, E. Zimanyi, "Hierarchies in a multidimensional model: From conceptual modeling to logical representation," Data & Knowledge Engineering 59 (2006) 348–377.

[16] R. Strohm. Oracle Database Concepts 11g. Oracle, Redwood City, CA 94065. 2007.

[17] I. Claudia and N. Galemmo, Mastering Data Warehouse Design -Relational and Dimensional. John Wiley and Sons, 2003, ISBN: 978-0-471-32421-8.

[18] Chuck Ballard, Dirk Herreman, Don Schau, Rhonda Bell, Eunsaeng Kim and Ann Valencic, Data Modelling Techniques for Data Warehousing, IBM, February 1998 : SG24-2238-00.